图书情报工作
第 65 卷 第 22 期 2021 年 11 月

基于 BERT-LDA 的关键技术识别方法及其实证研究*

——以农业机器人为例

■ 王秀红[1,2] 高敏[1]
[1] 江苏大学科技信息研究所 镇江 212013 [2]江苏大学图书馆 镇江 212013

摘 要：［目的/意义］好的关键技术识别方法能够更好地为各层各级的关键技术识别、预测和研发提供支撑。［方法/过程］提出基于 BERT-LDA 模型的关键技术识别方法，通过将 BERT 与 LDA 相结合，以弥补单一使用 LDA 主题模型缺乏上下文语义信息的缺陷，并以农业机器人为例进行实证研究。具体包括以下过程：①基于 python 构建 BERT 语义特征向量和 LDA 主题特征向量，将其在高维空间进行向量拼接，利用自编码器学习连接向量的低维潜在空间表示；②在潜在空间表示上使用 K-means 算法实现语义关联聚类，得到二维聚类效果图及关键技术主题词云图；③进行关键技术判定；④在农业机器人技术领域，与基于德温特 TI 专利软件的专利分析结果和《中国制造 2025》重点领域技术路线图中农业装备关键共性技术清单对比，实证本方法的有效性。［结果/结论］研究表明：BERT-LDA 模型提高了主题聚类的连贯性及细粒度划分的精准度；具有很好的关键技术识别精准率和召回率；对识别的不同数据库和出版类型的文献数据集具有较好的包容性与兼容性，适应性强；可广泛应用于各类关键技术的识别。

关键词：关键技术识别 农业机器人 BERT-LDA 模型 德温特专利
分类号：G251.2
DOI：10.13266/j.issn.0252－3116.2021.22.012

1 引言

21 世纪以来，全球科技创新进入空前密集活跃期，新一轮科技革命和产业变革正在重构全球创新版图、重塑全球经济结构。科学技术深刻影响着国家前途命运、人民生活福祉。习近平总书记在两院院士大会中提出“以关键共性技术、前沿引领技术、现代工程技术、颠覆性技术创新为突破口，敢于走前人没走过的路，努力实现关键核心技术自主可控，把创新主动权、发展主动权牢牢掌握在自己手中。”当前，我国科技创新在视野格局、创新能力、资源配置、体制政策等方面的短板日渐突显，关键核心技术受制于人的局面没有得到根本性改变。在新一轮的科技革命中，欲把握大势、抢占先机、大力发展关键核心技术、努力成为世界主要科学中心和创新高地，针对关键技术的识别与预测的相关研究变得尤为重要。

国内外众多学者开展了关键技术识别与预测研究，并取得了一定的研究成果。基于指标评估、专利数据、文本挖掘等方法的技术识别与预测研究，在一定程度上为世界各国的科技创新指引发展道路。随着科技情报需求的深化，对关键技术识别技术的创新优化提出了更严格的要求。鉴于此，笔者提出一种基于 BERT-LDA 的关键技术识别方法，以期提高主题的连贯性及细粒度划分的精准度；在确保关键技术识别的召回率和精准率的基础上，以增强关键技术识别对文献出版类型的包容性，不只如 TI（Thomson Innovation）一样局限于从专利文献中进行关键技术识别，仍可适用于不同数据库和出版类型的同语种科技文献摘要进行关键技术识别，必要时可将其整合，兼顾客观性、时效性。

2 相关文献回顾

国内外对关键新兴技术、共性技术、核心技术、突

* 本文系国家重点研发计划项目“农业装备制造产业集聚区域网络协同制造集成技术研究与应用示范”（项目编号：SQ2020YFB170242）研究成果之一。
作者简介：王秀红，研究馆员，博士，E-mail：xiuhongwang@ujs.edu.cn；高敏，硕士研究生。
收稿日期：2021－05－18 修回日期：2021－08－19 本文起止页码：114－125 本文责任编辑：徐健

破性技术等的关键技术识别研究已具有一定的基础，主要分为基于指标评估、基于专利网络和基于文本挖掘的三大类关键技术识别方法。

2.1 基于指标评估的识别方法

基于指标评估的识别方法是通过梳理关键技术的定义及特征，构建多指标评估的研究框架来识别关键技术。

其中，基于指标评估的关键技术识别方法较早受到学者的关注。例如：S. Altuntas 等通过综合考虑技术生命周期、扩散速度、专利权和扩展潜力 4 个指标评估相关技术[1]；I. Park 等根据专利的增长潜力、影响力和可销售性计算前景指数，并用于识别用户界面和用户体验技术领域的核心专利[2]；C. Lee 等提出使用多项专利指标的机器学习方法，用于识别早期阶段的新兴技术[3]；X. Liu 等在整合持久性、社区性和增长性的基础上提出三维评估框架来系统评估新兴技术[4]；江娴等通过构建基于专利分析的共性技术识别框架，从技术影响范围与技术研究阶段两个层面，以及基础性、外部性、集成性、超前性 4 个维度，识别共性技术[5]；杨武等提出基于专利数据利用指标体系探索核心技术的识别方法[6]；宋欣娜等通过构建新颖性、持久性、社区性和增长性的识别指标体系，并引入新兴分数和 LDA 主题模型，分别得到新兴术语和新兴主题，使用指标验证法验证其识别效果[7]。

基于指标评估的识别方法的关键是通过对关键技术的特征进行系统分析，整合专利数据的多项指标，形成关键技术指标评估体系，具有一定的科学性和有效性。部分指标的评估规则需经由专家根据技术特征设定，且需要专家参与部分指标的评分，识别结果过多依赖不同专家的不同认知和评判，识别结果的客观性有待提升。

2.2 基于专利网络的识别方法

专利网络分析是将社会网络理论与专利分析相结合，借助引用、共引、耦合等关联算法，对技术领域的演化网络及知识流网络进行分析的一种方法，主要是基于专利引用网络构建关键技术的识别框架。

例如：T. S. Cho 等通过分析专利引用网络，根据美国专利局在 1997－2008 年间授予台湾的专利，识别出 5 项核心技术和新兴技术[8]；M. H. C. Ho 等基于引文关系构建专利引文网络，通过路径分析识别出多次被引的核心专利[9]；O. Kuusi 等利用专利耦合网络预测纳米技术领域的突破性技术[10]；H. You 等提出基于专利和专利子类间知识转移的两层引文网络模型，并对技术发展趋势进行预测的方法，通过对相干光发生器分类专利进行实证研究，识别具有更大发展潜力的关键性技术[11]；李蓓等基于专利引用耦合聚类构建新兴技术识别模型及其相关指标体系，并对纳米技术领域的新兴技术进行识别[12]；杨艳萍等基于专利共被引聚类和专利组合分析构建关键技术识别分析框架[13]。

基于专利网络识别方法的关键是通过专利文献之间的引用关系构建引用网络，识别技术领域的关键技术。该方法在一定程度上避免了专家主观认知差异对识别结果的影响，可较客观地识别关键技术，但过多地依赖于实际的专利引文数据。一项专利从申请到公开再到授权有一定的周期，施引专利再经过申请、公开、授权又需要一定的时间，这使得专利文献之间的引证关系存在一定的滞后性，致使学者质疑基于专利网络及其引用特征进行关键技术识别结果的有效性和准确性。

2.3 基于文本挖掘的识别方法

基于文本挖掘的关键技术识别方法是基于论文和专利文献等文本内容，通过文本聚类、SAO 结构、LDA (Latent Dirichlet Allocation) 主题模型等自然语言处理技术方法挖掘深层次的技术隐性知识。

文本挖掘方法逐渐受到学者的重视，是目前具有最好识别效果的方法。H. Chen 等利用主题模型生成主题年份权重矩阵和基于主题的趋势系数序列，定量估计各个关键技术主题的发展趋势，并评价其对整个领域专利活动的贡献程度[14]；C. Yang 等利用半监督主题聚类模型整合技术领域知识，对 3D 打印行业技术分析，通过区分新主题和传统主题识别新兴技术[15]；Y. Zhou 等提出一种融合数据增强和深度学习方法的新方法，以克服深度学习在预测新兴技术时缺乏训练样本的问题[16]；李欣等利用文本挖掘抽取专利权利要求项中的 SAO 结构，基于改进的语义相似度算法对专利文本进行聚类，结合专利地图和语义分析识别新兴技术[17]；周源等提出一种基于机器学习主题模型的新兴技术识别方法，通过对技术领域全样本的论文与专利数据的高通量融合处理，挖掘论文与专利的语义信息，从而提高技术识别的全面性与颗粒度一致性[18]；陈伟等建立基于专利文献分析的关键共性技术识别框架，运用文本挖掘和技术演化分析方法，获取特定领域的关键共性技术[19]。基于文本挖掘的识别方法关键是利用数据挖掘和文本分析等对文本内容进行共现、聚类分析，能更客观、精准地识别关键技术。该类现有

的研究方法往往存在关键词之间缺少语义关联,忽略词语的上下文语义,难以抽取确切的关键技术主题,识别结果可解释性弱的问题。

综上所述,关键技术识别的现有研究方法多样、成果丰硕。基于指标评估和专利网络的技术识别方法,大多需要相关领域专家的参与,借由专家的专业知识和经验对技术的定性分析来开展技术识别工作,在识别过程中存在主观性强、时效性差、成本高等问题。基于文本挖掘的识别方法通过多源文本数据的分析实现关键技术识别的目的,该类方法具有可重复性强、成本高、客观准确的优势,但仍存在缺乏语义关联、识别结果可解释性弱的问题,需要有效体现语义关联提高识别结果的解释性的关键技术识别方法。

2.4 BERT 模型

近年来,自然语言处理模型在文本语义分析上取得了较好效果。Google 公司在 2018 年推出由 Devlin Jacob 等创建并发布的 BERT(Bidirectional Encoder Representations from Transformers)模型。BERT 模型是一种深度双向的、无监督的语言表示,是一种仅使用纯文本语料库进行预训练的模型[20]。而传统的模型是预先训练的、单向的、从左到右的。双向的好处是,该模型可以更好地学习语词之间的关系,检测语言的细微差别。

BERT 模型在文档语义研究、中文分词、词性标注、命名体识别、主题抽取等领域应用广泛,在主题抽取方面的应用已经取得一定进展。M. Asgari-Chenaghlu 等基于社交网络数据,使用 BERT 提供不同语境信息的语义关系,后借助 NoSQL、MongoDB 和 Neo4j 工具增强主题结果的可视化效果,实现社交媒体话题的实时检测[21];L. Thompson 等使用 BERT 结合聚类生成的主题与 LDA 主题模型相比较,结果表明 BERT 结合聚类效果更好[22];A. Abuzayed 等使用 BERTopic 使用不同的预先训练的阿拉伯语模型作为嵌入,并将其结果与 LDA 和 NMF(Non-negative matrix factorization)技术进行比较,结果表明 AraBERT 具有更好的性能[23];付静等针对短文本字数受限导致的特征稀疏和语义模糊的问题,提出一种基于 BERT-LDA 的新闻短文本分类方法[24];庄穆妮等为实现主题细粒度的舆情情感演化仿真,将 LDA 主题模型与 BERT 词向量深度融合,优化主题向量助力文本主题聚类[25];李越等提出了一种融合主题及上下文特征的汉缅双语词汇抽取方法,有效利用了汉缅双语主题的特征信息和上下文信息,进而抽取到质量更高的双语词汇[26]。

LDA 是一种非监督机器学习技术,可用来识别大规模文档集或语料库中潜藏的主题信息。采用词袋方法,将每一篇文档视为一个词频向量,从而将文本信息转化为了易于建模的数字信息[27]。但词袋法未考虑词与词之间的顺序,对复杂问题的学习效果较差,为模型的改进提供了契机。文献[21-23]等已有研究成果均证实 BERT 模型在主题抽取中的应用研究取得了更好的效果,主要得益于其特征表示能够表征上下文语义信息,可解决单词歧义、缺失语义表达能力等问题,因而经过预训练的 BERT 嵌入能够生成更有意义和更连贯的主题。BERT 结合聚类的模型简单、可靠,性能与 LDA 主题模型一样,甚至更好,即使在主题数量相对于数据集合规模较大的情况下,也能保持较高的主题质量。

已有学者将 BERT 模型与 LDA 主题模型相结合,应用于情感分析、文本分类、机器翻译等领域的研究,鲜有利用 BERT 模型结合 LDA 主题模型对主题聚类效果及其结果作进一步实证并验证的研究。在技术识别的研究领域,运用 LDA 主题模型识别关键技术主题的研究颇多,缺乏为弥补 LDA 主题模型难以表征上下文语义信息、单词歧义、缺失语义表达能力等缺陷的相关研究,致力于优化技术识别的效果。笔者从文本内容和语义关联关系出发,融合机器学习方法,提出基于 BERT-LDA 模型的关键技术识别的一般方法,并以农业机器人领域的专利文献为数据集进行实证研究,使用主题连贯性和轮廓系数将识别效果与 LDA、TF-IDF、Word2Vec 和 BERT 模型相比较,并与 TI 文本挖掘方法对比,包括专利地图与文本聚类,以验证 BERT-LDA 模型关键技术主题识别结果的有效性。进一步将本研究识别结果与《中国制造 2025》重点领域技术路线图中农业装备关键共性技术清单对比,验证关键技术判定结果的准确性。本研究的目的在于提供一种关键技术主题识别的一般方法,并提高识别结果的客观性、识别结果可解释性和精准率;不只如 TI 一样能从专利文献中识别关键技术,同样适用于从其他数据库和出版类型的科技文献中识别关键技术,且可以对同语种、不同出版形式的科技文献的技术描述文本整合后进行统一识别,从而大大提高关键技术识别的召回率和精准率,提升其适用性。

3 理论与方法

利用 BERT-LDA 模型识别关键技术主要包括以下过程:数据集构建与预处理、BERT-LDA 文本向量化表

示、语义关联聚类及其可视化和关键技术判定 4 个系统流程,具体思路如图 1 所示：

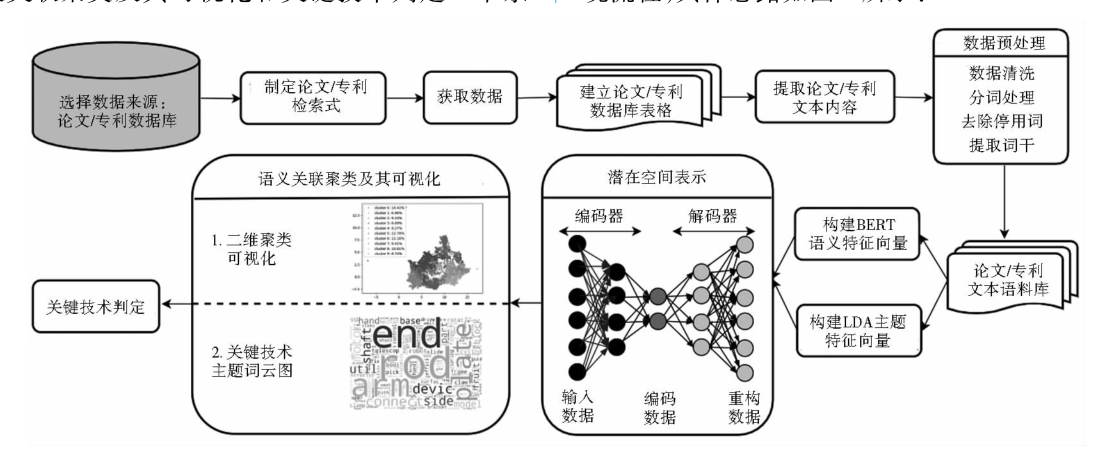


图 1　基于 BERT-LDA 模型的关键技术识别流程

3.1　数据集构建与预处理

确定检索使用的数据库,根据目标构建检索式,获取相应技术领域的文献。利用数据库提供的记录导出方式,提取标题、摘要、专利号、国际专利分类号和发明人等关键信息。对文本内容进行预处理,包括分词处理、停用词处理和词干提取等。笔者使用 Python 中的 NLTK 自然语言处理包以及 stop words 包对数据进行预处理。

3.2　BERT-LDA 文本向量化表示

笔者在大规模无标注数据集上训练 LDA 主题特征向量及 BERT 语义特征向量,再融合生成 BERT-LDA 文本向量化表示。

(1)构建 BERT 语义特征向量。利用 BERT 模型对预处理后的数据进行词嵌入,构建 BERT 语义特征向量。在 Transformer 编码器单元中,利用多头自注意力机制(Multi-Head Attention)处理后得到向量,经过残差连接和归一化层,再通过一个前馈网络和残差网络,提取到 BERT 语义特征向量。

将分词后的文档 di 输入模型,每个词被映射成 3 个向量;设定 ω、σ、ρ 分别为 BERT 模型获取文本的词向量、文本向量和位置向量,BERT 语义特征向量训练时,输出任意词语的 N 维向量表现形式。将 BERT 语义特征向量 d_m 定义为:

$$d_m = w_{ij}(\omega + \sigma + \rho) \qquad \text{公式 (1)}$$

(2)构建 LDA 主题特征向量。LDA 是一种包含词、主题和文档的三层贝叶斯概率模型,其降维思想为:将一篇分词后的文档降维为一个主题分布(如 n_0 个特征向量主题),根据对应的特征向量中的相关主题概率(n_0 个主题的概率相加为 1 即为主题分布)得到对应的文档主题。

假设文档由若干主题组成,则主题是由语料库中的所有特征词构成。设文本集 D 由 M 个文档组成,文档 d_i 包含 S 个句子,由 N_i 个词组成,w 表示词,z 为 w 所属的主题,每个词对应一个潜在主题。公式定义如下:

$$D = \{d_i \mid i \in \{1,2,\cdots,M\}\} \qquad \text{公式(2)}$$

$$d_i = \{d_i s \mid i \in \{1,2,\cdots,S\}\} \qquad \text{公式(3)}$$

$$w_i = \{w_{ij} \mid j \in \{1,2,\cdots,N_i\}\} \qquad \text{公式(4)}$$

$$z_i = \{z_{ij} \mid j \in \{1,2,\cdots,N_i\}\} \qquad \text{公式(5)}$$

LDA 主题模型的联合分布定义为:

$$P(w_i, z_i, \theta_i, \Phi \mid \alpha, \beta) = \prod_{j=1}^{N_i} P(w_{ij} \mid \varphi_{z_{ij}})\, P(z_{ij} \mid \theta_i) \cdot P(\theta_i \mid \alpha) \cdot P(\Phi \mid \beta) \qquad \text{公式(6)}$$

α 是每篇文档主题先验分布的超参数,θ_i 为参数 α 的 Dirichlet 分布采样;β 是每个主题内特征词先验分布的超参数,Φ 为参数 β 的 Dirichlet 分布采样。随后利用吉布斯抽样算法进行参数估计,迭代抽样直到收敛。模型训练结束输出语料库任意文本的主题分布矩阵,其向量维度与 BERT 语义特征向量维度相同。主题特征向量 μ 由每个主题的高频词与文档的余弦距离计算而得。

(3)向量拼接。由此,Transformer 编码器学习并存储了文档 d_i 的语义关系和语法结构信息,采用向量拼接的方式,将 BERT 语义特征向量与 LDA 主题特征向量叠加一起,形成新的输入向量,既包含词义特征又包含整体语义特征,定义为 d_m':

$$d_m' = \{\mu; d_m\} \qquad \text{公式(7)}$$

d_m' 表示融合 BERT 语义特征向量与 LDA 主题特征向量的文本向量化表示,“;”为向量拼接符号。

3.3　语义关联聚类及其可视化

由于向量拼接于信息稀疏的高维空间,本研究利

用自编码器学习连接向量的低维潜在空间表示,得到具有浓缩信息的低维表示。在潜在空间表示上使用聚类算法实现语义关联聚类,并从聚类中获得上下文主题。聚类的目的是将语义和主题上相似的语词分配到单个聚类中。最小平方和聚类算法(minimum square sum clustering, MSSC)最适合于大数据聚类,最具代表性的是用于聚类的 K-means 算法。K-means 算法是一种高效的聚类算法,有距离平方和(SSD)的目标函数可估计得到的聚类质量,其只有一个参数 K 即所需的簇数。参数 K 等价于一个确定主题建模中主题数量的参数,因此从主题建模到 K-means 有一个自然的联系。K-means 具有以下特点:简单、高效、参数数量最少,不需要初步计算距离矩阵,可能进行大数据处理等优点,这使得 K-means 在解决各种 NLP 任务时成为对上下文信息的嵌入进行聚类的最佳选择,故本研究采用 K-means 算法。

聚类时需提前确定聚类簇数量,即 K 值,笔者利用困惑度(perplexity)确定最优主题数。困惑度随着主题数量的增加而递减,当曲线趋于平缓时的主题数即可作为最优的主题数量。将每个主题下概率排名较高的主题词进行可视化分析,并输出排名前 10 的主题词及其概率值,从而进一步判定关键技术。

通过构建 BERT-LDA 关键技术识别模型,能够充分结合上下文语义信息,弥补 LDA 主题模型的劣势,训练出更优的主题向量,得到具有更好细粒度和聚类精准度的关键技术识别效果。

4 实证研究

以专利文献为例进行数据收集与处理,借助国际权威的德温特专利数据和 TI 专利分析软件的强大主题识别功能,将本研究的关键技术识别结果分别与 TI 专利分析的主题识别结果和《中国制造 2025》重点领域技术路线图中农业装备关键共性技术清单进行比对,验证 BERT-LDA 模型关键技术识别的精准率和召回率。

4.1 数据收集及预处理

选取德温特专利数据库中农业机器人领域专利作为数据样本,在文献调研和专家知识的基础上,最终确定农业机器人领域专利的检索式:((TS = (agricultur * or crop or crops or fruit or fruits or vegetable * or harvest * or seedling *))or (MAN = (X25 – N * or X22 – X11 or X22 – P09 or Q19 – G or T06 – D01 * or A12 – W04 * or X25 – X02 *)) or (IP = (A01B * or A01C * or A01D * or A01F * or A01G * or A01M – 021 *))) AND ((TS = (robot * or manipulator * or "mechanical arm" or "mechanical arms" or "mechanical hand" or "mechanical hands")) or IP = (B25J *) or (MAN = (X25 – A03E * or T06 – D07B * or V03 – U14 * or V04 – M30R * or V04 – Q30R * or V06 – U05 * or V04 – R04F1 * or X27 – U * or S05 – B07 *)) not (IP = (A01G – 005 * or A01G – 023 *)) or (MAN = (X25 – N02 * or T06 – D01C)))。检索数据覆盖范围为 2020 年 12 月 6 日前公开的所有农业机器人专利文献。对数据进行处理和筛选后共获取专利 8 957 件,提取专利号/申请号、DWPI 标题、DWPI 摘要、国际专利分类号 IPC、申请日等相关信息,完成数据收集。

对数据进行预处理,具体过程如下:对 DWPI 摘要文本进行数据清洗,去除摘要缺失的共获取专利 8 912 件;进行分词,过滤标点与数字,同时进行去噪处理,主要包括:小写转化、拼写检查更正、单复数统一、同义词合并、全称和缩写,去除停用词(如 a,for 等)、专有描述词(如 comprise,involve 等)、学术词汇(如 novelty,use,advantage 等)、出现频率高但对具体关键技术识别结果没有意义的领域高频干扰词(如 robot, agriculture 等),以及提取词干等数据预处理操作。

4.2 BERT-LDA 模型识别结果与分析

为增强 BERT 模型对本研究问题的适应性,基于 GOOGLE 的 BERT 基本预训练模型,利用农业机器人专利摘要语料库对预训练的 BERT 语言模型进行微调,其中向量嵌入维度是 768 维,得到改进后的 BERT 预训练模型。随后利用改进的 BERT 预训练模型和 LDA 主题模型对清洗后的 DWPI 摘要训练向量并将其叠加拼接,紧接着聚类生成农业机器人领域的关键技术主题。利用困惑度的变化估计最优的农业机器人关键技术主题数量。困惑度随主题数量的变化情况如图 2 所示。当主题数取 10 时,BERT-LDA 模型的困惑度值趋于稳定,故选择主题数为 K = 10。

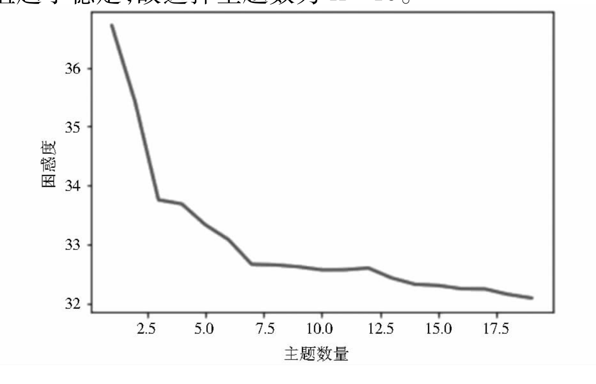


图 2 困惑度随主题数量的变化情况

数据降维可视化最新工具 UMAP（Uniform Manifold Approximation and Projection）在可视化质量方面保留了更多的全局结构，具有优越的运行性能和可扩展性[28]。采用 UMAP 降维工具对 BERT-LDA 模型识别出的关键技术主题进行可视化，聚类结果如图 3 所示。识别出的 10 个主题分布明确，类内具有较高的连贯性与一致性，表明 BERT-LDA 模型的聚类效果颇佳。

在 BERT-LDA 模型识别出农业机器人技术领域的 10 个关键技术主题的基础上，选取每个主题下概率 TOP50 主题词进行可视化分析，来确定关键技术主题内容，绘制的 10 个关键技术主题词云图如图 4 所示。对应的 BERT-LDA 模型下每个主题排名前 10 的特征词及其概率分布见表 1。

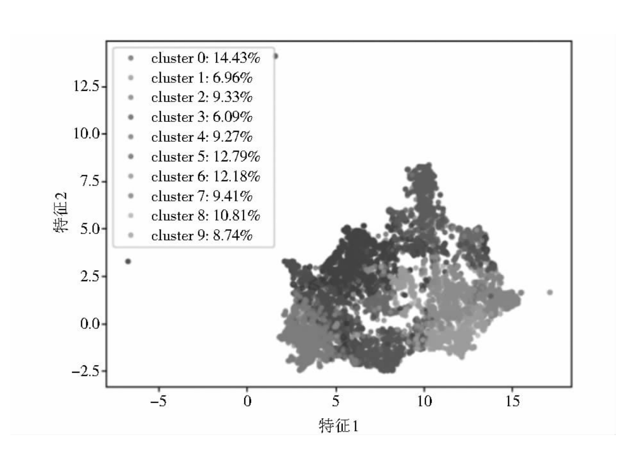


图 3　基于 UMAP 的二维聚类可视化









图 4　BERT-LDA 模型下农业机器人的关键技术主题词云图

基于 BERT-LDA 模型识别出 Topic1-Topic10 关键技术主题，为农业机器人技术领域中出现概率较高的特征词构成的集合，每个主题 Topic 均可视为该领域中的一个研究热点。从表 1 各主题下的特征词即可知该领域的研究热点。

Topic1 中的 TOP10 主题词为 connect、pick、collect、arm、automatic 等特征词。结合农业机器人技术领域的国际专利分类号及德温特手工代码，明确该主题对应农业机器人领域的"用于采摘的自动装置"关键技术。通过对检索结果进行文本分析，进一步验证其准确度。如江苏大学的 CN101273688-A 专利，为橘子采摘机器人设计的柔性采摘装置，具体采用双眼立体视觉系统连接主计算机和工作控制机，通过运动控制卡控制机械臂和终端执行器的关节。在该专利的题名与摘要中出现了 pick、arm 与 motor 等特征词，与本研究的识别结果一致。如日本井关农机的 JP2008206438-A、中国农业大学的 CN101356877-A 专利、西北农林科技大学的 CN202232196-U 等专利。

Topic2 中的 TOP10 主题词为 position、signal、direction、distance 等特征词，同样的方法对应为领域的"目标的位置探测与定位"关键技术。如瑞典胡斯华纳的 EP3346348-A1 专利，用于引导机器人园艺工具的方法，

表 1 BERT-LDA 模型下农业机器人的关键技术特征词及其概率统计

主题编号	属于该主题的高概率特征词
Topic1	connect(0.140 2), fruit(0.128 7), pick(0.091), harvest (0.074 8), arm (0.063 2), mechanism(0.061 6), automatic(0.059 2), control (0.055), collect(0.053 7), motor(0.052 9)
Topic2	sensor(0.239 6), position(0.103 4), device(0.084 6), signal(0.083 6), direction(0.065 4), navigation (0.061 7), boundary(0.058 8), distance(0.046), detect(0.045 3), data(0.038 6)
Topic3	vehicle (0.126 2), control(0.124 9), unit(0.102 9), autonomous(0.090 6), drive(0.063), wheel(0.061 7), direction(0.059 4), motor (0.055 8), path(0.055 1), position(0.055 1)
Topic4	method(0.131), vehicle(0.101 1), autonomous(0.078 5), area(0.075 7), path(0.069 3),unmanned(0.067 1), navigate(0.064 1), mobile(0.059 7), signal(0.058 8), system(0.057 3)
Topic5	seedling(0.144 9), mechanism(0.114), fix(0.100 5), transplant(0.095 5), arm(0.077 8),manipulate(0.069 8), plant(0.054 1), convey (0.052), end(0.051 5), transmission (0.045 1)
Topic6	control(0.149 1), connect(0.093 3), finger(0.078 1), module(0.071 3), end(0.069 5),device(0.067 2), remote(0.060 3), manipulator (0.058 5), sensor(0.055 6), unit(0.055)
Topic7	graft(0.125 4), cut(0.118), plant(0.106 1), transplant(0.088 9), part(0.085 8),hand(0.057 8), hypocotyl(0.054 8), supply(0.053 5), seedling(0.051 5), position(0.050 7)
Topic8	lawn(0.180 5), mower(0.097 6), method(0.089 1), unit(0.087 1), system(0.064 3),area(0.063 3), sensor(0.056 4), signal(0.053 5), lawnmower(0.046 5), tool(0.044 8)
Topic9	water(0.148 8), irrigation(0.092 1), storage(0.089 4), box(0.082 1), device(0.074 1),plant(0.073 5), pipe(0.057 4), tank(0.056 2), automatic(0.051 6), fruit(0.048 2)
Topic10	cut(0.134 9), pruning(0.108 5), rod(0.098), trim(0.077 5), machine(0.075),end(0.074 2), clip(0.061 4), tree(0.056 4), plate(0.054 3), branch(0.046 6)

注：主题内容栏中的结构为特征词(概率分布)

在机器人工具引导系统中使用的机器割草机预定位置，具体涉及由机器人园艺工具在不同距离的引导线跟踪磁信号。在该专利的题名与摘要中出现了 position、signal 与 distance 等特征词，与本研究的识别结果一致。本领域综合实力排名靠前的美国约翰迪尔、LG 电子公司、日本洋马等在农业机器人的"目标的位置探测与定位"技术领域也有一定的站位，如美国约翰迪尔的 US2011295424-A1 专利、LG 电子公司的 KR2015125508-A 专利、日本洋马的 JP2020119595-A 专利等。

Topic3 中的 TOP10 主题词为 vehicle、control、drive、direction 等特征词，同样的方法对应为领域的"转向控制"关键技术。如日本井关农机的 JP2020166534-A 专利，是农业机器人拖拉机等工作车辆，其控制器配置为当转弯路线连接的直线前进路线之间的距离超过预定距离时，选择两轮驱动模式。该专利在题名与摘要中出现了 vehicle、control 与 drive 等特征词，与本研究的识别结果一致。本领域综合实力 TOP 排名专利权人德国博世、日本洋马、美国约翰迪尔等在"转向控制"技术领域的专利占有重要份额，如德国博世的 DE102007023157-A1 专利、日本洋马的 JP2019061695-A 专利、美国约翰迪尔的 US2012085458-A1 专利等。

Topic4 中的 TOP10 主题词为 method、autonomous、path、navigate 等特征词，同样的方法对应为领域的"自主导航与路径规划"关键技术。如瑞典胡斯华纳的 SE201650022-A1 专利，是自航机器人工具导航方法，根据第一、二信号路径的分离距离和路径长度差，计算从基站到机器人工具的方位代表值。该专利在题名与摘要中出现了 method、navigate 与 path 等特征词，与本研究的识别结果一致。农业机器人领域综合实力 TOP 排名的德国博世、IROBOT 公司、美国约翰迪尔等在农业机器人的"自主导航与路径规划"技术领域的专利占有重要份额，如德国博世的 DE102011003064-A1 专利、IROBOT 公司的 US2018116105-A1 专利、美国约翰迪尔的 US2010094499-A1 专利等。

同理可推出 Topic5-Topic10 的主题分别对应农业机器人领域的"种苗的移栽机械""机械手的控制装置""嫁接""割草机""灌木装置和修剪""整枝或立木打枝工具"，具体关键技术主题名称如表 2 所示：

表 2 农业机器人的关键技术主题名称

主题	关键技术主题名称
Topic1	用于采摘的自动装置
Topic2	目标的位置探测与定位
Topic3	转向控制
Topic4	自动导航与路径规划
Topic5	用于种苗的移栽机械
Topic6	机械手的控制装置
Topic7	嫁接
Topic8	割草机
Topic9	灌木装置
Topic10	修剪、整枝或立木打枝工具

4.3 关键技术判定

利用BERT-LDA模型识别出的农业机器人TOP三大关键技术为:末端执行器、目标的探测与定位技术和自动导航与路径规划技术。

(1)末端执行器。综合topic1、topic5、topic6、topic7和topic10中的特征词,可知"末端执行器"技术是农业机器人领域的关键技术之一。末端执行器一般由机械装置和传感器组成,主要包括机器人抓手、碰撞传感器、旋转连接器、压力工具等,作用于采摘、移栽、喷雾等农业生产作业过程。由于农业机器人作业环境和目标具有复杂性和特殊性,末端执行器的设计需充分考虑其特性,以保证作业质量。需要重视末端执行器的创新设计,提升其通用性、精确性、灵活性及可控性。

(2)目标的探测与定位技术。整合topic2、topic5和topic7中的特征词,可知"目标的探测与定位"技术是农业机器人领域的关键技术之一。农业机器人对作业目标的精准识别及定位是其开展作业的前提,目前主要采用机器视觉技术,它是人类研究较早的一种环境感知技术,最早起源于美国。由于作业过程中光照条件、目标遮挡、个体差异等问题的存在,目标的探测与定位技术仍需进一步发展与完善。在未来的研究中,通过将机器视觉技术与其他技术相融合,改进图像获取和图像处理算法,提高目标探测及定位的准确性和精准度。

(3)自动导航与路径规划技术。整合topic3、topic4、topic8和topic9中的特征词,可知"自动导航与路径规划"技术是农业机器人的关键技术之一。农业机器人的自主行动需要导航系统的指挥,根据感知的环境信息和目标位置,做出行动路径规划,并在无人干涉的情形下,自主移动到预定的位置,目前主要采用视觉导航和以其为主的组合导航方法。农业机器人在执行作业过程中,由于作业环境的复杂性、作业目标分布的随机性及动态情况的不可预知性等问题,对自动导航与路径规划提出了更严格的要求。

4.4 识别结果检验

笔者利用主题建模的连贯性和聚类的轮廓系数检验BERT-LDA模型的识别效果。主题连贯性(CV Coherence)基于滑动窗口,对主题词进行one-set分割(一个set内的任意两个词组成词对进行对比),并使用归一化点态互信息和余弦相似度间接获得连贯度,用于衡量同一主题内的特征词语义是否连贯,其取值范围为[0,1]。轮廓系数(Silhouette Score)是测量类内一致性的指标,用于评价模型聚类效果的好坏,其取值范围为[-1,1]。连贯度和轮廓系数数值越高意味着模型效果越好。基于农业机器人领域德温特专利摘要数据集,对比5种不同方法的主题模型的关键技术识别情况,结果如表3所示。进一步对比二维聚类可视化效果,结果如图5所示。

表3 5种主题建模方法的关键技术识别效果对比

模型评估系数	LDA	TF-IDF	Word2Vec	BERT	BERT-LDA
主题连贯性	0.458	0.478	0.481	0.453	0.508
轮廓系数	/	0.006	0.071	0.054	0.150

由表3和图5可知,BERT-LDA模型的主题连贯性数值为0.508,其他4种方法的数值均低于0.5,可见BERT-LDA模型的识别结果在同一主题内的特征词具有更好的连贯性,有效提高了识别结果的可解释性。由轮廓系数及二维聚类可视化图比较模型的聚类效果,BERT-LDA模型的轮廓系数为0.15,二维聚类可视化图中显现出各主题类间划分明确,反观其他3种方法轮廓系数的最高数值仅有0.071,各主题类间多有重合,难以辨析各主题类间的边缘。对比结果表明,基于BERT-LDA模型的关键技术识别的聚类效果明显提高。

德温特创新平台TI是全球权威可靠的专利数据和专利分析平台,具备强大的智能检索、分析、预警和海量文献图像化功能,协助组织建立跨部门的专利技术情报搜集与分析能力,为用户提供更广泛视角的技术信息来源。它收录了全球156个国家/地区的专利信息,涵盖全球75个国家/地区的专利全文,并通过人工的方法把收集的不同语种的专利文献的摘要统一用英文并用自然语言进行改写,避免了专利文献检索分析的跨语言障碍和专利描述语的晦涩难懂,这使得TI成为目前全球权威高端的专利技术识别工具。由于德温特创新平台是商业数据库,其文本挖掘的后台算法属商业机密,无法公开获取,无法从其算法层面进行描述对比分析,只能从识别结果进行对比分析。因此,笔者将实证结果与基于德温特TI专利分析软件的专利分析结果进行比对,以检验笔者提出的BERT-LDA模型识别关键技术方法的可行性与有效性,以及识别的精准率和召回率。

专利地图将专利文献以图像化的方式表现出技术主题的全景,其中TI专利分析中的文本聚类将专利文献自动分类成多个聚类簇,输出文本聚类列表。基于上述检索式,绘制农业机器人领域的TI专利地图如图6所示,图中山峰海拔高度代表特定主题文献的密度大小。在同样的数据检索范围内,将TI专利分析的主题文本聚类结果与BERT-LDA模型识别结果进行对比,结果见表4。

TF-IDF: cluster 0: 0.48%, cluster 1: 3.10%, cluster 2: 24.43%, cluster 3: 9.40%, cluster 4: 7.65%, cluster 5: 18.82%, cluster 6: 2.49%, cluster 7: 12.91%, cluster 8: 5.43%, cluster 9: 12.28%

Word2Vec: cluster 0: 12.35%, cluster 1: 6.73%, cluster 2: 10.94%, cluster 3: 13.05%, cluster 4: 11.69%, cluster 5: 11.91%, cluster 6: 10.12%, cluster 7: 8.59%, cluster 8: 6.12%, cluster 9: 8.49%

BERT: cluster 0: 14.96%, cluster 1: 4.57%, cluster 2: 7.52%, cluster 3: 9.04%, cluster 4: 10.90%, cluster 5: 11.79%, cluster 6: 13.90%, cluster 7: 10.42%, cluster 8: 10.64%, cluster 9: 6.28%

BERT-LDA: cluster 0: 14.43%, cluster 1: 6.96%, cluster 2: 9.33%, cluster 3: 6.09%, cluster 4: 9.27%, cluster 5: 12.79%, cluster 6: 12.18%, cluster 7: 9.41%, cluster 8: 10.81%, cluster 9: 8.74%

（坐标轴：特征 1，特征 2）

图 5　4 种模型 UMAP 二维聚类可视化效果对比图

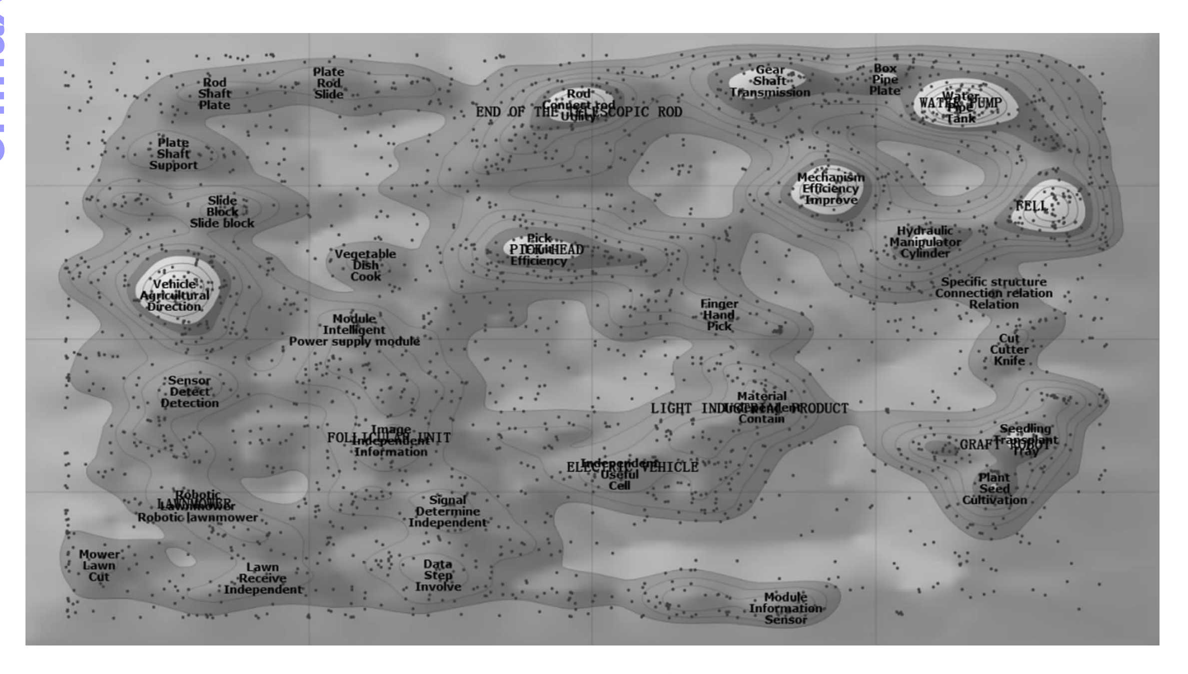


图 6　农业机器人领域专利地图

王秀红，高敏. 基于 BERT-LDA 的关键技术识别方法及其实证研究——以农业机器人为例[J]. 图书情报工作,2021,65(22):114－125.

表 4　BERT-LDA 模型与 TI 文本聚类下农业机器人关键技术主题内容对比

序号	BERT-LDA		TI 文本聚类	
	主题内容	主题特征词	主题内容	主题特征词
1	用于采摘的自动装置	connect, fruit, pick, harvest, arm, mechanism, automatic, control, collect, motor	果蔬采摘装置	pick, fruit, mechanical, vegetable, system, platform, robot, sort, belt, convey
2	目标的位置探测与定位	sensor, position, device, signal, direction, navigation, data, boundary, distance, detect,	目标的位置探测与定位	object, target, location, signal, point, position, image, detection, data, mobile
3	自动导航与路径规划	method, vehicle, autonomousarea, path, unmanned, navigate, mobile, signal, system	自动导航与路径规划	information, vehicle, work, navigate, path, data, device, autonomous, mobile, line
4	用于种苗的移栽机械	seedling, mechanism, fix, transplant, arm, manipulate, plant, convey, end, transmission	播种、移栽装置	seedling, finger, transplant, fruit, pick, graft, scion, stock, plug, transplant, plant
5	机械手的控制装置	control, connect, finger, module, end, device, remote, manipulator, sensor, unit	机械手的控制装置	end, rod, connect rod, module, system, control, pick, fruit, mechanical, remote
6	嫁接	graft, cut, plant, transplant, part, hand, hypocotyl, supply, seedling, position	嫁接	cut, cutter, tap, graft, rubber, glue, tool, barrier, tree, harvest
7	割草机	lawn, mower, method, unit, system, area, sensor, signal, lawnmower, tool	割草机	robotic, mow, mower, tool, work, mow robot, vehicle, mobile, lawnmower, autonomous
8	灌木装置	water, irrigation, storage, box, device, plant, pipe, tank, automatic, fruit	灌木装置	water, tank, pipe, irrigation, rod, plate, arm, mechanical, system, area
9	修剪、整枝或立木打枝工具	cut, pruning, rod, trim, machine, end, clip, tree, plate, branch	修剪、整枝工具	prune, arm, hydraulic, climb, tree, robot, pair, connect rod, garden, trim, green
10	转向控制	vehicle, control, unit, drive, autonomous, wheel, direction, motor, path, position	液体喷雾设备	spray, pesticide, medicine, plate, slide, stir, arm, mechanical, tree, plant
	关键技术主题识别结果内容一致性		90%	

结合图 6 和表 4 可知,TI 专利分析结果显示:农业机器人领域的 Top10 关键技术主题归纳为果蔬采摘装置、目标的位置探测与定位、自动导航与路径规划、播种、移栽装置、机械手的控制装置、嫁接、割草机、灌木装置、修剪、整枝工具和液体喷雾设备,与笔者提出的方法识别结果对比,关键技术主题识别结果的内容一致性高达 90%,充分证实了 BERT-LDA 模型识别关键技术的有效性。根据表 4 进一步深入对比 BERT-LDA 模型与 TI 专利分析文本聚类的主题特征词,由于 BERT-LDA 模型考虑了文本的语义和上下文信息,识别出的各主题下的特征词表现出具有更好的语义连贯性,提高了识别结果的可解释性,确保了 BERT-LDA 模型在关键技术识别时具有高的精准率和召回率。

对比国家制造强国建设战略咨询委员会组织编制的重点领域的技术路线图即"《中国制造 2025》重点领域技术路线图",其中涉及农业装备的机器人末端执行器、可靠性试验方法、检测控制技术、传感器等相关技术,进一步验证了本研究关键共性技术识别结果与实际情况的吻合性。

5　研究结论与展望

笔者提出的基于 BERT-LDA 模型的关键技术识别方法,其有效性在主题连贯性、轮廓系数及二维聚类可视化效果方面都得到验证。以农业机器人技术领域专利数据为例进行实证,通过与 TI 的专利地图和文本聚类方法以及《中国制造 2025》重点领域技术路线图中农业装备关键共性技术清单进行对比分析,验证 BERT-LDA 模型识别关键技术的精准率和召回率。同时,克服了 TI 专利分析软件的主题识别方法只限应用于德温特专利文献、不适于也不能同时应用于非专利文献的技术主题分析的缺陷。

5.1　研究结论

实证结果表明:与现有的 LDA、TF-IDF、Word2Vec、BERT 模型相比,BERT-LDA 模型可充分考虑文本的语义信息和上下文信息,与 LDA 主题模型相融合,可明显提高关键技术识别时主题的连贯性及细粒度划分的精准度。与国际权威的 TI 专利分析主题聚类结果对比,BERT-LDA 模型在识别关键技术时同样具备很好的识别精准率和召回率。经本方法进行关键共性技术判定得到末端执行器、目标的探测与定位、自动导航与路径规划技术,与《中国制造 2025》重点领域技术路线图中农业装备的关键共性技术清单结果相比较,识别结果相一致,验证了本研究关键共性技术识别结果的准确性。

BERT-LDA 模型用于关键技术识别时,不仅能适应专利文献,同样适用于期刊论文、会议文献、学位论

文、研究报告等不同出版类型的技术文献。需要时可将同语种的不同数据库的文献进行整合,如将 WOS、EBSCO、Science Direct、IEEE 等不同数据库中的技术文献的摘要进行整合后分析,利用 BERT-LDA 模型在技术领域的全部技术出版物中进行统一检索和关键技术识别,在确保识别精准率的前提下可大大提高关键技术识别的召回率。与现有的模型相比,BERT-LDA 模型在关键技术识别时具有较好的包容性与兼容性,适应性强。

5.2 研究展望

为了与国际权威的 TI 分析结果进行比对,笔者选择了与 TI 相同的数据集进行训练和实证,数据来源于德温特数据库中专利文献 DWPI 摘要文本。缺少对期刊论文、会议论文、研究报告等多源科技文献的整合,在未来的研究中,将利用 BERT-LDA 模型整合不同数据库、不同出版类型的文献数据集进行研究,以实现更全面的关键共性技术识别。为提高主题分析及文本聚类的效果,将在数据收集及预处理环节做更多的优化,包括数据清洗、停用词扩充、词干提取等。为进一步提高 BERT-LDA 模型关键技术识别结果的可解读性,可考虑引入 SAO 结构,以将语词之间的关系具体化为某个技术方面的"问题"和"解决方案"及其之间的对应关系,进一步提高模型的识别结果的可解释性。后续研究中也将考虑在此基础上进一步结合专家调查与指标评估等方法对关键共性技术的判定做进一步的改善研究。

参考文献:

[1] ALTUNTAS S, DERELI T, KUSIAK A. Forecasting technology success based on patent data[J]. Technological forecasting and social change, 2015, 96(7): 202-214.

[2] PARK I, PARK G, YOON B, et al. Exploring promising technology in ICT sector using patent network and promising index based on patent information[J]. ETRI journal, 2016, 38(2): 405-415.

[3] LEE C, KWON O, KIM M, et al. Early identification of emerging technologies: a machine learning approach using multiple patent indicators[J]. Technological forecasting and social change, 2018, 127(2): 291-303.

[4] LIU X, PORTER A L. A 3-dimensional analysis for evaluating technology emergence indicators[J]. Scientometrics, 2020, 124(1): 27-55.

[5] 江娴, 魏凤. 基于专利分析的共性技术识别研究框架[J]. 情报杂志, 2015, 34(12):79-84.

[6] 杨武, 杨大飞. 基于专利数据的产业核心技术识别研究——以5G 移动通信产业为例[J]. 情报杂志, 2019, 38(3):39-45, 52.

[7] 宋欣娜, 郭颖, 席笑文. 基于专利文献的多指标新兴技术识别研究[J]. 情报杂志, 2020, 39(6):76-81,88.

[8] CHO T S, SHIH H Y. Patent citation network analysis of core and emerging technologies in Taiwan: 1997-2008[J]. Scientometrics, 2011, 89(3): 795-811.

[9] HO M H C, LIN V H, LIU J S. Exploring knowledge diffusion among nations: A study of core technologies in fuel cells[J]. Scientometrics, 2014, 100(1): 149-171.

[10] KUUSI O, MEYER M. Anticipating technological breakthroughs: using bibliographic coupling to explore the nanotubes paradigm [J]. Scientometrics, 2007, 70(3): 759-777.

[11] YOU H, LI M, HIPEL K W, et al. Development trend forecasting for coherent light generator technology based on patent citation network analysis[J]. Scientometrics, 2017, 111(1): 297-315.

[12] 李蓓, 陈向东. 基于专利引用耦合聚类的纳米领域新兴技术识别[J]. 情报杂志, 2015, 34(5):35-40.

[13] 杨艳萍, 董瑜, 韩涛. 基于专利共被引聚类和组合分析的产业关键技术识别方法研究——以作物育种技术为例[J]. 图书情报工作, 2016, 60(19):143-148,124.

[14] CHEN H, ZHANG G, ZHU D, et al. Topic-based technological forecasting based on patent data: a case study of Australian patents from 2000 to 2014 [J]. Technological forecasting and social change, 2017, 119(6): 39-52.

[15] YANG C, ZHU D, WANG X, et al. Requirement-oriented core technological components' identification based on SAO analysis [J]. Scientometrics, 2017, 112(3): 1229-1248.

[16] ZHOU Y, DONG F, LIU Y, et al. Forecasting emerging technologies using data augmentation and deep learning[J]. Scientometrics, 2020, 123(1): 1-29.

[17] 李欣, 王静静, 杨梓, 等. 基于 SAO 结构语义分析的新兴技术识别研究[J]. 情报杂志, 2016, 35(3):80-84.

[18] 周源, 刘宇飞, 薛澜. 一种基于机器学习的新兴技术识别方法:以机器人技术为例[J]. 情报学报, 2018, 37(9):939-955.

[19] 陈伟, 林超然, 孔令凯,等. 基于专利文献挖掘的关键共性技术识别研究[J]. 情报理论与实践, 2020, 43(2):92-99.

[20] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint, 2018, arXiv:1810.04805.

[21] ASGARI-CHENAGHLU M, FEIZI-DERAKHSHI M R, FARZINVASH L, et al. TopicBERT: a cognitive approach for topic detection from multimodal post stream using BERT and memory-graph [J]. Chaos, solitons & fractals, 2021, 151(10): 111274.

[22] THOMPSON L, MIMNO D. Topic modeling with contextualized word representation clusters[J]. arXiv preprint, 2020, arXiv: 2010.12626.

[23] ABUZAYED A, AL-KHALIFA H. BERT for Arabic topic modeling: An experimental study on BERTopic technique[J]. Procedia computer science, 2021, 189(11): 191-194.

[24] 付静,龚永罡,廉小亲,等. 基于 BERT-LDA 的新闻短文本分类方法[J]. 信息技术与信息化,2021(2):127－129.
[25] 庄穆妮,李勇,谭旭,等. 基于 BERT-LDA 模型的新冠肺炎疫情网络舆情演化仿真[J]. 系统仿真学报,2021,33(1):24－36.
[26] 李越,毛存礼,余正涛,等. 融合主题及上下文特征的汉缅双语词汇抽取方法[J]. 小型微型计算机系统,2021,42(1):91－95.
[27] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation [J]. The journal of machine learning research, 2003, 3(1): 993－1022.
[28] MCINNES L, HEALY J, MELVILLE J. Umap: uniform manifold approximation and projection for dimension reduction[J]. arXiv preprint, 2018, arXiv:1802.03426.

作者贡献说明:
王秀红:论文选题与设计,研究思路设计、实验设计,指导论文写作、修改、定稿;
高敏:负责论文框架设计,完成数据采集、处理及论文撰写。

The Key Technology Identification Method Based on BERT-LDA and Its Empirical Research: A Case Study of Agricultural Robots

Wang Xiuhong[1,2] Gao Min[1]
[1] Institute of Science and Technology Information, Jiangsu University, Zhenjiang 212013
[2] Jiangsu University Library, Zhenjiang 212013

Abstract: [**Purpose/significance**] A good key technology identification method can provide better support for key technology identification, prediction and research and development at all levels. [**Method/process**] In this paper, a key technology identification method based on Bert-LDA was proposed, which combined BERT and LDA to make up for the lack of contextual semantic information in a single LDA topic model. An empirical study was carried out with agricultural robots as an example. Specifically, it included the following processes: ① Constructing BERT semantic feature vector and LDA topic feature vector based on Python, combining them in a high-dimensional space, and learning the low-dimensional latent space representation of the concatenated vector by using an autoencoder; ② In the potential space representation, K-means algorithm was used to realize semantic association clustering, and the effect diagram of two-dimensional clustering and key technology subject word cloud maps were drawn; ③ Determining key technologies; ④ In the field of agricultural robots, the effectiveness of this method was demonstrated by comparing with the results of TI patent analysis and the list of key generic technologies for agricultural equipments in the "Made in China 2025" technology roadmap for key areas. [**Result/conclusion**] The results show that the Bert-LDA model improves the coherence of topic clustering and the accuracy of fine-grained classification. With a good key technology identification accuracy and recall rate, there are good inclusiveness, compatibility and adaptability to the identified literature data sets of different databases and publishing types. It can be widely used to identify all kinds of key technologies.

Keywords: key technology identification agricultural robots BERT-LDA model Derwent patents